Unicode, UTF-8, ASCII, and SNOMED CT®

John Kilbourne, MD[1], Tim Williams[1]

[1]College of American Pathologists, Northfield, IL

ABSTRACT

SNOMED CT text files are encoded using UTF-8 to allow worldwide distribution and use of the terminology. Incorporating such UTF-8 encoded text into a system not currently using UTF-8 is simplified when the specific range of characters in the imported data is known. This poster describes the superset of ASCII found in the SNOMED CT US/UK Edition January 2003 release.

INTRODUCTION

UTF-8 is one of the Unicode Transformation Formats which convert a Unicode "codepoint" or hexadecimal integer into a particular sequence of bytes[1]. The Unicode standard maps code points to a set of characters (including diacritical marks, ligatures and other glyphs), for the purpose of standardizing the computer representation of the writing systems of the world. UTF-16 and UTF-32 are other transformation formats; the numbers "8", "16" and "32" refer to the number of bits per unit or byte. UTF-8 encodes Unicode characters into a sequence of 8-bit bytes. The standard has a capacity for over a million distinct codepoints and is a superset of all characters in widespread use today. By comparison, ASCII (American Standard Code for Information Interchange) includes 128 character codes. Eight-bit extensions of ASCII, (such as the commonly used Windows-ANSI codepage 1252 or ISO 8859-1 "Latin-1") contain a maximum of 256 characters.[2] Each 8-bit extension to ASCII differs from the rest. For characters represented by the 7-bit ASCII character codes, the UTF-8 representation is exactly equivalent to ASCII, allowing transparent round trip migration. Other Unicode characters are represented in UTF-8 by sequences of up to 6 bytes, though most Western European characters require only 2 bytes[3].

PROBLEM ADDRESSED

Each SNOMED CT concept is associated with a number of human readable strings ("descriptions")[4]; some of those incorporate characters lying outside the range of ASCII. For instance, Löffler's syndrome, Ménière's syndrome, and Lieberkühn's crypts are among the SNOMED descriptions that cannot be represented properly using ASCII. The conversion from UTF-8 to the various 8-bit ASCII extensions is not wholly transparent. The translation to ISO 8859-1 "Latin-1" is algorithmic; other translations require a table of substitution values. Identifying the specific superset of ASCII used in SNOMED CT would simplify its migration into an existing system.

RESULTS

All Unicode characters with a codepoint value greater than U+7F (decimal 128) in the January 2003 SNOMED CT Descriptions Table (US/UK edition) were identified, and 28 distinct non-ASCII characters in 406 separate descriptions were found. All codepoints had a value less than U+FF (decimal 256), meaning all characters are contained within ISO-8859-1 "Latin-1".

DISCUSSION

An international medical terminology encoded in the international "alphabet" (Unicode and its UTF implementations) encourages adoption by users around the world. The UTF-8 character encoding in SNOMED CT allows both for SNOMED's future use with any standard writing system in the world (including ideographs), and the relatively painless adoption into existing information systems currently using only Western European characters.

CONCLUSIONS

Incorporating SNOMED CT US/UK edition into an existing information system that uses an 8-bit extension of ASCII should pose relatively few problems, as all characters lie within the range of ISO 8859-1, a very commonly used superset of ASCII.

REFERENCES

1. http://www.unicode.org
2. http://czyborra.com/charsets/codepages.html
3. http://czyborra.com/utf/#UTF-8
4. SNOMED Clinical Terms Technical Specification; http://www.snomed.org

Special thanks to John M. Dluglosz